Global Sensitivity Analysis using Polynomial Chaos Expansion on the Grassmann Manifold

Valentina Bazyleva, Victoria M. Garibay[0000-0003-0399-0591], and Debraj Roy[0000-0003-1963-0056]

Faculty of Science, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands.
valya.bazyleva@student.uva.nl, {v.m.garibay, d.roy}@uva.nl

Abstract. Traditional global sensitivity analysis (GSA) techniques, such as variance- and density-based approaches, are limited in cases where a comprehensive understanding of temporal dynamics is critical, especially for models with diverse timescales and structural complexity, such as system dynamics and agent-based models (ABMs). To address this, we propose a novel manifold learning-based method for GSA in systems exhibiting complex spatiotemporal processes. Our method employs Grassmannian diffusion maps to reduce the dimensionality of the data and polynomial chaos expansion (PCE) to map stochastic input parameters to diffusion coordinates of the reduced space. We calculate sensitivity indices from PCE coefficients, aggregating multiple outputs and their entire trajectories for a more general estimation of parameter sensitivities. We demonstrate the capabilities of the proposed approach by applying it to the Lotka-Volterra model and an epidemic dynamics ABM and capturing diverse temporal dynamics. We establish that the new methodology meets all "good" properties of a global sensitivity measure, making it a valuable alternative to traditional GSA techniques. We anticipate that it will potentially expand the application of manifold-based approaches and deepen the understanding of complex spatiotemporal processes.

Keywords: Global sensitivity analysis · Sobol' indices · Agent-based modelling · Grassmannian diffusion maps · Polynomial chaos expansion.

1 Introduction

Parametric global sensitivity analysis (GSA) is an essential tool for enhancing model efficiency. The determination of which parameters and combinations thereof contribute the most to model uncertainty can allow for the development of simplified models. This directs the focus of experiments toward the parameters of greatest influence. Eliminating or fixing parameters at a certain value can also provide a substantial computational advantage. When developing a computationally intensive agent-based model (ABM), reducing model complexity is of great interest. Frequently in ABMs, heterogeneous, simulated populations interact, make decisions, and take action at every time step, so increasing the speed of these calculations has a cumulative advantage.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

The current standard for GSA relies on the use of either Sobol' or density-based methods. A critical disadvantage of using these methods is the inability to aggregate the results of calculations for each individual time step into a single index or, alternatively, the loss of information when only considering the final outcome of the model in the analysis, ignoring the progression at individual time steps. ABMs are characteristically stochastic and often subject to non-linear interaction effects. Due to the frequent appearance of multimodal and fat-tailed distributions in ABM results, conducting a GSA presents a unique challenge [21]; sensitivity is not well captured with traditional Sobol' analyses as the technique is variance-based. While it is possible to employ density-based methods in cases with poorly defined variance, the described issues in aggregation still apply.

To address these limitations, the proposed analysis method inclusively considers the outputs of interest at each time step. Then, the resulting high-dimensional data is organised into a tensor and projected onto a Grasmann manifold, with the goal of detecting that the data can be mapped to a subspace of lower dimension. The reduced-dimension data is used in conducting a sensitivity analysis with polynomial chaos expansion (PCE) methods. PCE, used in mapping numerical model input to output, becomes difficult to apply to cases with high dimensionality. However, assuming sparse effects makes it possible to create representative surrogate models from fewer samples, lowering the computational expense of PCE. The next sections begin with the context under which the methods in this study were developed, including a guided review of related work. Then, an overview of the proposed GSA method is provided, followed by demonstrative applications to a classic Lotka-Volterra dynamical system and a large-scale ABM of disease dynamics. These applications culminate in an assessment and discussion of the performance of the novel methodology.

2 Related Work

Parametric variance-based GSA, specifically with Sobol'/Saltelli ANOVA techniques, has become widely adopted for ABMs across various fields due to its robust sensitivity estimates for non-linear models with parameter interactions [2,23,29,33]. However, a crucial limitation of these methods is that variance in model outputs is not always attributed to uncertainty [21], and sensitivity assessment in ABMs for verification and validation is often insufficiently explored [22]. This could be attributed to the lack of tools and methodologies focusing on a comprehensive analysis of ABM dynamics. One such inefficient use of Sobol'/Saltelli variance-based GSA is estimating sensitivity based solely on the final time step, disregarding the preceding trajectory. This issue has been addressed by time-dependent GSA, which calculates Sobol' sensitivity indices at multiple time steps throughout the simulation [16].

The computational cost of the Sobol'/Saltelli method based on estimating high-dimensional integrals via crude Monte-Carlo (MC) simulation is a practical concern for researchers due to slow error convergence [29]. More efficient sampling schemes have been proposed, such as Latin hypercube sampling and low-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

discrepancy or Sobol' sequences [19,28], along with direct formulas for evaluating Sobol' indices with fewer model evaluations [15,20,24,27,32]. However, calculating Sobol' indices for complex and large-scale spatial ABMs remains challenging due to computational costs, particularly in models with systemic variability from aleatory uncertainty [1]. Averaging over multiple ABM repetitions has been suggested to address this issue [18], with some researchers leveraging multi-GPU parallel computing and high-performance computing resources for GSA of large-scale ABMs [31].

Besides MC and quasi-Monte-Carlo (QMC) methods, stochastic polynomial methods such as PCE can be used for constructing surrogate models to approximate ANOVA decomposition [5,30]. The calculation of Sobol' indices for PCE emulators is analytical, reducing the computational cost to that of computing PCE coefficients [30]. Despite PCE's advantages and the growing research on sparse methods tackling the issue of exponential increase in the number of polynomial basis functions with increasing input dimension—resulting in an excessive computational cost for models with high-dimensional input, also known as "curse of dimensionality" in uncertainty quantification (UQ) literature [8,13,17]—PCE-based ANOVA decomposition is not commonly applied for GSA in ABMs with only few examples of successful applications [4,8,10].

Sparse PCE methods help reduce the size of basis functions and experimental design required for GSA but can become computationally intractable for high-dimensional output [13]. Dimensionality reduction techniques, including linear and non-linear methods, can address the challenge of model fidelity for UQ tasks when, instead of a scalar or low-dimensional vector, high-dimensional responses are of interest [6,25]. While linear methods can extract the dominant modes of data, they are ineffective in capturing the non-linear geometries of a dataset. Transcending the limitation of linear techniques, non-linear methods for dimensionality reduction posit that high-dimensional data resides on a manifold, which is a low-dimensional and more informative space [6,25]. Thus, kernel-based diffusion maps (DMaps) can discover low-dimensional manifolds embedded in Euclidean space and be exploited to construct accurate, lower-cost surrogate models [11,12]. However, exceedingly high-dimensional data, such as numerical simulations with many degrees of freedom, may not be well-described in Euclidean space and may inherently reside on a submanifold of a Riemannian manifold [6].

A proposed extension from and complement to DMaps is Grassmannian diffusion maps (GDMaps) addressing limitations when dealing with data exhibiting geometric structures on a Riemannian submanifold [6,25]. This is achieved by combining pointwise linear dimensionality reduction with a multipoint, non-linear dimensionality reduction step using DMaps with a suitable Grassmannian kernel. GDMaps are particularly fitting for high-dimensional data represented by vectors or matrices, where Euclidean metrics cannot meaningfully describe distances between objects, thus capturing the geometric structures spanning the data [6]. Dos Santos et al. present a simple example illustrating GDMaps' capability to capture intrinsic geometric structures in data [6]. This example

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

demonstrates that while conventional DMaps can find a low-dimensional representation, the resulting manifold may not accurately represent the underlying subspace structure of the data (see SI Section A).

Leveraging the suitability of GDMaps for latent representation of very high-dimensional data on a lower-dimensional manifold, Kontolati et al. [14] proposed a surrogate model construction method capable of generating out-of-sample predictions from a limited number of observations. The "encoder" path of the technique combining GDMaps and PCE for dimensionality reduction and mapping between input parameters and diffusion coordinates provides a framework for statistical moment estimation from PCE coefficients in the latent space [14], potentially enabling sensitivity index calculations from PCE coefficients. However, to the authors' knowledge, GSA methods employing GDMaps and PCE have not yet been suggested.

3 Methods

3.1 Variance-based Global Sensitivity Analysis: Sobol' Indices

We consider a set of d independent random variables (RVs) $\boldsymbol{X} = \{X_i\}_{i=1}^d$, serving as an input into a model $Y = f(\cdot)$. For simplicity, we assume that the RVs X_i are uniformly distributed on $[0,1]$: $Q_i \sim \mathcal{U}(0,1)$, $\Gamma = [0,1]^d$ and write the Sobol' decomposition of the response $f(\mathbf{X})$ as the finite, hierarchical expansion:

$$\begin{aligned} f(\boldsymbol{X}) &= f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{i,j\neq i}^{d} f_{ij}(X_i, X_j) + \cdots + f_{12\ldots d}(\boldsymbol{X}) \\ &= f_0 + \sum_{\boldsymbol{u}\subset\{1,\ldots,d\}} f_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}), \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{\boldsymbol{u}} := \{X_{i_1}, \ldots, X_{i_s}\}$ and the summands satisfy the orthogonality condition: $\int_\Gamma f_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}) f_{\boldsymbol{v}}(\boldsymbol{X}_{\boldsymbol{v}}) d\boldsymbol{X} = 0 \quad \forall \boldsymbol{u} \neq \boldsymbol{v}$. In Eq. (1), f_0 is the mean response of f, the univariate functions $f_i(X_i)$ quantify independent contribution given the individual parameters, the bivariate functions $f_{ij}(X_i, X_j)$ represent the interactions of X_i and X_j on the response with similar interpretations for higher-order interaction effects [26].

From the total variance theorem, the total variance $\mathbb{V}[Y] = D$ can be decomposed as $D = \sum_{i=1}^d D_i + \sum_{1\leq i<j\leq d} D_{ij} + \cdots + D_{12\ldots d}$, which we use to define first- and total-order Sobol' indices as

$$S_i = \frac{D_i}{D} = \frac{\mathbb{V}[\mathbb{E}(Y|X_i)]}{\mathbb{V}[Y]}, \quad S_{T_i} = 1 - \frac{D_{\sim i}}{D} = 1 - \frac{\mathbb{V}[\mathbb{E}(Y|X_{\sim i})]}{\mathbb{V}[Y]} = \frac{\mathbb{E}[\mathbb{V}(Y|X_{\sim i})]}{\mathbb{V}[Y]}. \tag{2}$$

Calculation of Sobol' Indices with Conventional Methods Using MC, we obtain the following estimators for mean as $\hat{f}_0 = 1/N \sum_{n=1}^N f(X^{(n)})$, and for total variance as $\hat{D} = 1/N \sum_{n=1}^N f^2(X^{(n)}) - \hat{f}_0^2$. To obtain the estimates for

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

D_i and $D_{\sim i}$, we use Saltelli's algorithm (explained in [24,26]) to reduce the number of evaluations from N^2 for crude MC to $N(d+2)$ by constructing three types of X samples: $X = (X_1, \ldots, X_d)^\top$, its complete resample $X' = (X'_1, \ldots, X'_d)^\top$, and $(X_i, X'_{\sim i}) = (X'_1, \ldots, X'_{i-1}, X_i, X'_{i+1}, \ldots, X'_d)^\top$, with $i = 1, \ldots, d$, where all factors except for X_i are resampled. Thus, the estimates for D_i and $D_{\sim i}$ become $\hat{D}_i = 1/N \sum_{n=1}^{N} f(X_i^{(n)} X_{\sim i}^{(n)}) f(X_i^{(n)} X'^{(n)}_{\sim i}) - \hat{f}_0^2$ and $\hat{D}_{\sim i} = 1/N \sum_{n=1}^{N} f(X_i^{(n)} X'^{(n)}_{\sim i}) f(X'^{(n)}_i X'^{(n)}_{\sim i}) - \hat{f}_0^2$, respectively. The derived estimators $\hat{D}$, $\hat{D}_i$ and $\hat{D}_{\sim i}$ are then used to calculate the first- and total-order Sobol' indices in Eq. (2).

Sobol' Indices Using PCE PCE describes the input-output relationship using polynomials orthogonal with respect to the probability density function (PDF) of the input RVs. Sobol' decomposition of a PCE results from reordering terms of the truncated PCE approximating $f(\boldsymbol{X})$, written as $\widetilde{\mathcal{E}}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \eta_{\boldsymbol{\alpha}} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{X})$, where $\mathcal{A}$ is a total-degree multi-index set, $\eta_{\mathbf{s}}$ are corresponding PCE coefficients, and $\Phi_{\mathbf{s}}(\mathbf{X})$ are multivariate orthonormal polynomials with respect to $f_{\mathbf{X}}$ such that $\langle \Phi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \Phi_{\boldsymbol{\beta}}(\boldsymbol{X}) \rangle = \int_Z \Phi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \Phi_{\boldsymbol{\beta}}(\boldsymbol{X}) f_{\boldsymbol{X}}(\boldsymbol{X}) d\boldsymbol{X} = \gamma_{\boldsymbol{\alpha}} \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$. We can obtain interaction sets as $\mathcal{A}_{\boldsymbol{u}} = \{\boldsymbol{\alpha} \in \mathcal{A} : t \in \boldsymbol{u} \Leftrightarrow \alpha_t \neq 0\}$ for a given $\boldsymbol{u} := \{i_1, \ldots, i_s\}$, leading to the following decomposition: $\widetilde{\mathcal{E}}(\boldsymbol{X}) = \mathcal{E}_0 + \sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \mathcal{E}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})$, with $\mathcal{E}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}) := \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\boldsymbol{u}}} \eta_{\boldsymbol{\alpha}} \Phi_{\boldsymbol{\alpha}}(\boldsymbol{X})$, resulting in the following general expression for PCE-based Sobol' indices, which can be derived analytically at any order from the PCE coefficients [30]:

$$S_{\boldsymbol{u}} = D_{\boldsymbol{u}}/D = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\boldsymbol{u}}} \eta_{\boldsymbol{\alpha}}^2 / \sum_{\boldsymbol{\alpha} \in \mathcal{A} \setminus \mathbf{0}} \eta_{\boldsymbol{\alpha}}^2. \tag{3}$$

3.2 GSA Using Grassmannian Diffusion Maps and PCE

The methodology presented in Algorithm 1 largely draws from GDMaps technique proposed by Dos Santos et al. [6] and manifold learning-based PCE developed by Kontolati et al. [14]. Refer to the corresponding papers for a thorough description of the Grassmann manifold principles and other elements of differential geometry essential for developing GDMaps. See SI Section B.1 for more details on the proposed methodology.

The GDMaps method is an extension of the conventional DMaps that consists of two stages: a linear pointwise dimension reduction and a non-linear multipoint dimension reduction. The implementation of GDMaps is outlined in Lines 1-5 of Algorithm 1. This first step projects each element of the data on the Grassmann manifold, or Grassmannian, denoted as $\mathcal{G}(p, n)$ and defined as a p-dimensional subspaces embedded in n-dimensional Euclidean space $\mathbb{R}^n$. The parameter p, which relates p-dimensional subspaces embedding, is closely tied to the notion of matrix rank or the number of linearly independent matrix columns. The non-linear dimensionality reduction step consists of building a valid kernel (Lines 3-4 in Algorithm 1) and running the DMaps algorithm (Line 5 in Algorithm 1). The projection kernel (see SI Section B.1) is adopted throughout this research. Using

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

Algorithm 1: GSA using PCE on the Grassmann manifold

Input: Experimental design $\mathscr{X} = \{\boldsymbol{X}_i \in \mathbb{R}^k\}_{i=1}^{\mathcal{N}}$ and model response concatenated into a single vector and reshaped into $n \times m$ matrix $\mathscr{Y} = \{\mathcal{M}(\boldsymbol{X}_i)\}_{i=1}^{\mathcal{N}} = \{\boldsymbol{Y}_i \in \mathbb{R}^{n\times m}\}_{i=1}^{\mathcal{N}}$; dimension of the Grassmann manifold p; a number of diffusion coordinates to retain g.

Output: First-order Sobol' indices $\{\boldsymbol{S}_i \in \mathbb{R}^g\}_{i=1}^k$; total-order Sobol' indices $\{\boldsymbol{S}_{T_i} \in \mathbb{R}^g\}_{i=1}^k$; approximated PCE error ϵ_{val}.

1 **for** $i \leftarrow 1$ **to** $\mathcal{N}$ **do**

2 Perform singular value decomposition (SVD): $\boldsymbol{Y}_i = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^T$, where $\boldsymbol{\Sigma}_i \in \mathbb{R}^{p\times p}$ is a diagonal matrix containing singular values. $\boldsymbol{U}_i \in \mathbb{R}^{n\times p}$ and $\boldsymbol{V}_i \in \mathbb{R}^{m\times p}$ are orthonormal matrices.

3 For each pair $[\boldsymbol{U}_i, \boldsymbol{U}_j]$ and $[\boldsymbol{V}_i, \boldsymbol{V}_j]$ compute the entries $k_{i,j}$ of the kernel matrices $k_{ij}(\boldsymbol{U})$ and $k_{ij}(\boldsymbol{V})$ using e.g., projection kernel as $k_{pr}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left\|\boldsymbol{X}_i^T \boldsymbol{X}_j\right\|_F^2$.

4 (Optionally) construct a composed Grassmannian diffusion kernel $K(\boldsymbol{U}, \boldsymbol{V})$ by taking the Hadamard product of the corresponding kernels: $k(\boldsymbol{U}, \boldsymbol{V}) = k_{i,j}(\boldsymbol{U}) \circ k_{i,j}(\boldsymbol{V})$, or by summing $k_{i,j}(\boldsymbol{U}) + k_{i,j}(\boldsymbol{V})$.

5 Run the DMaps with a Grassmannian kernel to obtain first g non-trivial diffusion coordinates $\{\boldsymbol{\Theta}_i \in \mathbb{R}^g\}_{i=1}^N$, and their respective eigenvectors $\{\psi_k\}_{k=1}^g$ with $\psi_k \in \mathbb{R}^{\mathcal{N}}$ and eigenvalues $\{\lambda_k\}_{k=1}^g$.

6 Construct a total-degree multi-index set Υ (with cardinality $\#\Upsilon = S$) that satisfy $\|\mathbf{s}\|_1 \leq s_{\max}, s_{\max} \in \mathbb{Z}_{\geq 0}$, leading to a PCE basis of size $\frac{(k+s_{\max})!}{k!s_{\max}!}$.

7 Construct PCE approximation $\widetilde{\mathcal{E}}(\mathbf{X}) = \sum_{\boldsymbol{s}\in\Upsilon} \eta_{\boldsymbol{s}} \boldsymbol{\Phi}_{\boldsymbol{s}}(\boldsymbol{X})$ where $\eta_{\boldsymbol{s}} \in \mathbb{R}^g$ is computed by solving the least square problem $\arg\min_{\zeta\in\mathbb{R}^{\#\Lambda}} \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left\{\mathcal{E}(\boldsymbol{X}_i) - \sum_{\boldsymbol{s}\in\Upsilon} \eta_{\boldsymbol{s}} \boldsymbol{\Phi}_{\boldsymbol{s}}(\boldsymbol{X}_i)\right\}^2$.

8 Approximate PCE generalisation error by calculating the validation error as $\epsilon_{\text{val}} = \frac{\sum_{i=1}^{\mathcal{N}_*} \left(\boldsymbol{\Theta}_i^* - \widetilde{\mathcal{E}}(\boldsymbol{X}_i^*)\right)^2}{\sum_{i=1}^{\mathcal{N}_*} \left(\boldsymbol{\Theta}_i^* - \bar{\boldsymbol{\Theta}}^*\right)^2}$, where $\{\boldsymbol{X}_i^* \in \mathbb{R}^k\}_{i=1}^{\mathcal{N}_*}$ and $\{\boldsymbol{\Theta}_i^* \in \mathbb{R}^q\}_{i=1}^{\mathcal{N}_*}$ comprise a test set, chosen to be of size $\mathcal{N}_* = \frac{1}{3}\mathcal{N}$; $\bar{\boldsymbol{\Theta}}^* = \frac{1}{\mathcal{N}_*} \sum_{i=1}^{\mathcal{N}_*} \boldsymbol{\Theta}_i^*$ is the mean response of the test set on the latent space.

9 Obtain first-order Sobol' indices $\{\mathbf{S}_i \in \mathbb{R}^g\}_{i=1}^k$ and total-order Sobol' indices $\{\mathbf{S}_{T_i} \in \mathbb{R}^g\}_{i=1}^k$ using Eqs. (8) and (9) from SI Section B.1, respectively.

GDMaps, we acquire g diffusion coordinates $\boldsymbol{\Theta}_i$ for the top g non-trivial eigenvalues (non-parsimonuous implementation). To address the issue of "repeated eigendirections" in complex data, we also utilise parsimonious representation employing local linear regression to identify unique eigendirections [7].

Next, we calculate Sobol' indices on the manifold using PCE following Kontolati et al.'s approach [14]. PCE is used to approximate mapping between input parameters and corresponding model responses projected on the latent space (i.e., coordinates on the diffusion manifold) $\mathcal{E} : \boldsymbol{X} \rightarrow \boldsymbol{\Theta}$ as $\widetilde{\mathcal{E}}(\boldsymbol{X}) = \sum_{\boldsymbol{s}\in\Upsilon} \eta_{\boldsymbol{s}} \boldsymbol{\Phi}_{\boldsymbol{s}}(\boldsymbol{X})$. The implementation of the approach employing the least square method to obtain vector-valued PCE coefficients $\eta_{\boldsymbol{s}} \in \mathbb{R}^g$ is outlined in Lines 6-8 of Algorithm 1, with Line 8 used to calculate validation error

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

to evaluate the surrogate's accuracy. From Section 3.1, Sobol' indices can be acquired without extra calculations by collecting multi-indices related to partial variance caused by individual random inputs (first-order effects) or combined with other random inputs (total-order effects) into two multi-index sets. As the PCE coefficients have a vector-valued dimension equal to the retained diffusion coefficients g, we estimate the first- and total-order Sobol' indices for each of the g diffusion coordinates.

3.3 Applications

A classic dynamical system (Lotka-Volterra [14]) and an ABM (DeepABM-COVID [3]) were selected as a sample demonstration of the range of model types for which the framework is applicable. These were utilised to compare, albeit indirectly, the performance of the proposed GSA framework and GSA employing conventional Sobol' index calculation methods over multiple time steps. The two models used in illustrating the application of the proposed framework and the setup used for the evaluation are described in SI Sections B.2 and B.3.

4 Results

Application 1: Lotka-Volterra Dynamical System First- and total-order sensitivity indices, along with 95% bootstrap confidence intervals, were computed for the Lotka-Volterra dynamical system with two uncertain parameters α and β at fifteen evenly spaced time steps (Figs. 1a to 1d). Oscillatory behaviour in both main and total-effect indices corresponds to the behaviour of the model outputs for the defined parameter ranges for α and β (see SI Fig. C.1). The difference between the resulting first- and total-order indices for both outputs is small, hence the variance in model output is predominantly due to main effects rather than interactions.

Mean and variance of first- and total-order Sobol' indices were obtained using the GSA framework with GDMaps PCE from fifty resampled input matrices (Figs. 1e and 1f). Three first- and total-order indices were derived from the PCE coefficients for each output. Three non-trivial, parsimoniously selected diffusion coordinates were used, converging to $\boldsymbol{\Theta}_i = \{\theta_1, \theta_2, \theta_5\}$ for one solution[1], which can be found as 2D plots in SI Fig. C.2. In Figs. 1e and 1f, both main and interaction effects of α and β have close values, with β slightly higher for the first two diffusion coordinates. Interaction effects are significantly larger for the second and third coordinates. While direct comparison with Figs. 1a to 1d is inappropriate due to different data representations, the proposed framework arguably better highlights parameter differences in terms of their influence on output variance. The new GSA approach reveals a more apparent distinction between main and total-effect indices compared to the conventional time-dependent GSA methods. A similar comparison for the Lotka-Volterra model with four uncertain parameters is presented in SI Fig. C.3.

[1] Different sets of diffusion coordinates are possible for each resampled solution.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

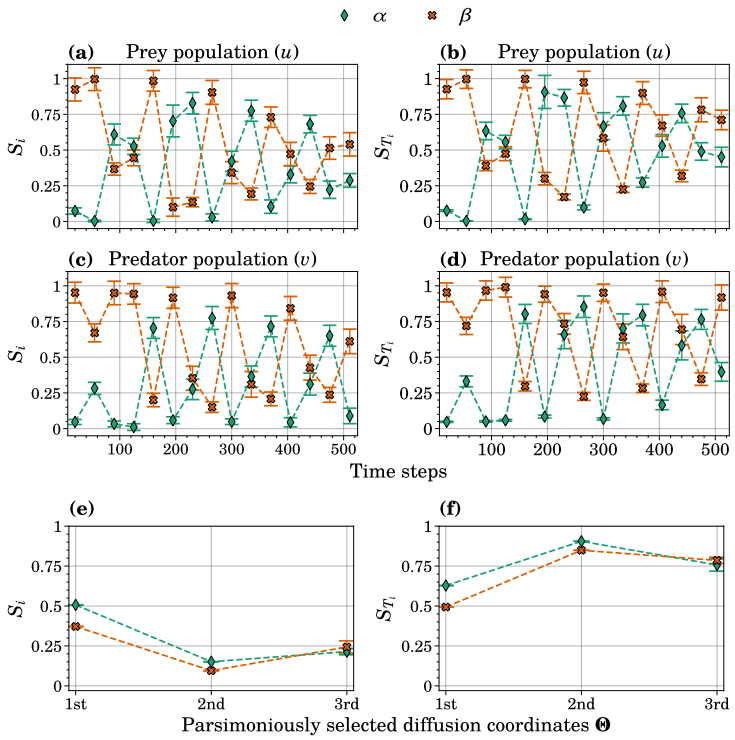

Fig. 1. Estimates of first- and total-order sensitivity indices, S_i and S_{T_i}, respectively, of α, and β (see SI Section B.2 Tab. 1) for two model output measures: number of prey per time step u, and number of predators per time step v, using (a-d) conventional Sobol' index calculation methods and (e, f) the GSA framework employing GDMaps PCE. Error bars indicate 95%-bootstrap confidence intervals in (a-d), and variance from fifty resamples in (e, f). For GDMaps PCE, Grassmannian dimension $p = 10$ and maximal polynomial degree $s_{\max} = 6$ were used.

Application 2: DeepABM-COVID We applied the proposed framework to estimate Sobol' indices on the Grassmann manifold using the DeepABM-COVID model outputs. Data generation and the general procedure for GSA framework are outlined in SI Section B.3. For GDMaps, we considered six dimensions, $p = \{3, 13, 17, 22, 23, 24\}$, corresponding to most frequently occurring ranks in the entire dataset (twenty runs). The frequency of occurrence and selection threshold are shown in Fig. 2a. Non-trivial eigenvalues[2] for chosen dimensions are presented in Fig. 2b. Given the physical interpretation of DMaps based on Markov Chain timescales, the regions around unstable equilibria (slow modes) with the largest eigenvalues correspond to the slowest possible ergodic dynamics in a system. From Fig. 2b, $p = 3$ is attributed to the slowest dynamics on the manifold

[2] The first zero-indexed eigenvalue, λ_0, is a trivial eigenvalue, which is always $\lambda_0 = 1$.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

compared to other dimensions. Smaller p values correspond to larger λ_i values, particularly for $i = 1$ and $i = 2$, due to lower p allowing less detailed data representation on the Grassmannian, resulting in a more coarse-grained subspace structure revealed by DMaps performed on the manifold.

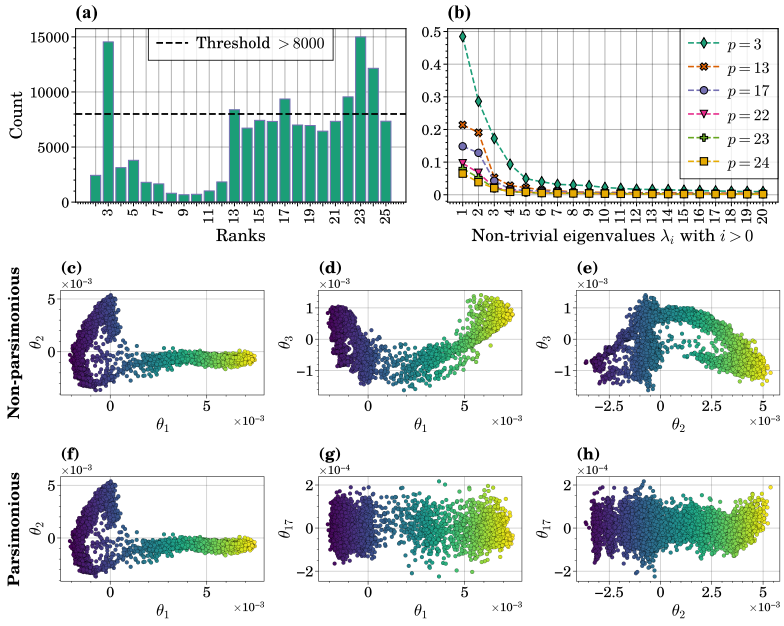

Fig. 2. (a) The frequency of rank occurrence in the entire data (twenty runs) with a threshold used for selection of the Grassmannian dimension, p. (b) Scree-plot of eigenvalues from GDMaps on DeepABM-COVID model output matrix $\mathscr{Y} \in \mathbb{R}^{7168\times900}$ for run 3 and six Grassmannian dimensions p. (c-h) 2D plots of three diffusion coordinates from GDMaps on DeepABM-COVID model output for run 16, using $p = 13$. Diffusion coordinates converged to $\boldsymbol{\Theta}_i = \{\theta_1, \theta_2, \theta_{17}\}$ for parsimonious representation (f-h).

We examined both parsimonious and non-parsimonious implementations to retain diffusion coordinates. Figs. 2c to 2h present 2D plots of retained diffusion coordinates ($g = 3$) for both implementations. The example demonstrates the case when the first two coordinates coincide, but the third parsimoniously selected one corresponds to shorter timescale dynamics indicated by the scale of the y-axis. The remaining 2D plots for other Grassmannian dimensions are in SI Figs. C.5 (non-parsimonious) and C.6 (parsimonious). Notably, for larger p, parsimonious representation selected diffusion coordinates with lower corresponding eigenvalues more frequently than for smaller p. This relates to Fig. 2b, where larger p values are attributed to lower initial eigenvalues and exhibit lower decay rates. Fig. 3 shows three examples of distributions of parsimoniously selected diffusion coordinates. Multimodal distributions correspond to higher eigenvalues,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

while unimodal and Gaussian-like distributions are attributed to lower eigenvalues, as in Fig. 3f. Additional examples can be found in SI Figs. C.7 and C.8.

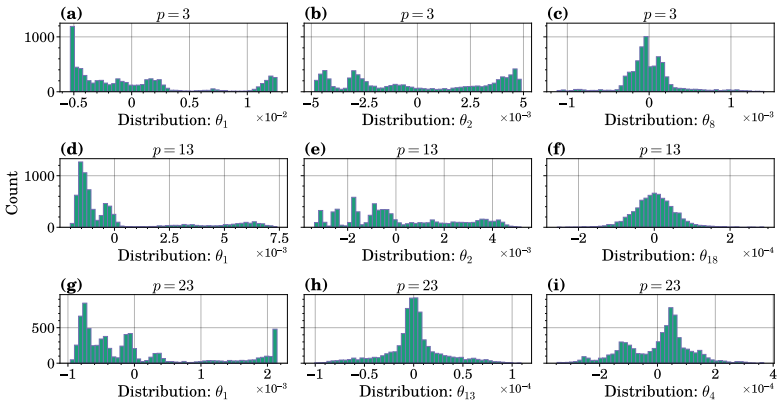

Fig. 3. Distributions of parsimoniously selected diffusion coordinates for the Grassmannian dimensions: $p = 3$ (a-c), $p = 13$ (d-f), and $p = 23$ (g-i) obtained from GDMaps run 17 of the DeepABM-COVID model output $\mathscr{Y} \in \mathbb{R}^{7168\times900}$. Higher values in the subscript correspond to lower non-trivial eigenvalues λ_i, ordered from high to low.

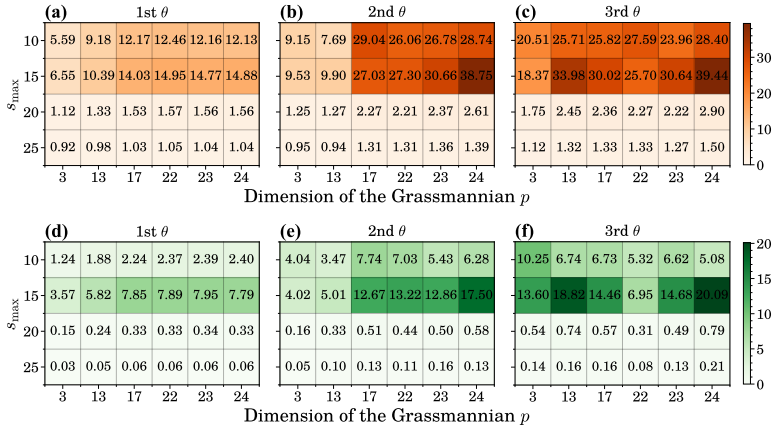

Fig. 4. Heatmaps of the mean (a-c) and standard deviation (d-f) of validation error from GDMaps PCE averaged over twenty runs for Grassmannian dimensions $p = \{3, 13, 17, 22, 23, 24\}$ and maximal polynomial degrees $s_{\max} = \{10, 15, 20, 25\}$. The colour maps are between zero and the maximal values rounded to the second decimal place.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

To construct PCE surrogates with the total-degree PCE basis, we evaluated different maximum polynomial degrees $s_{\max} = \{10, 15, 20, 25\}$. Validation errors averaged over twenty runs, and standard deviations are presented in Figs. 4a to 4c and Figs. 4d to 4f, respectively, for each Grassmannian dimension p, and maximum polynomial degree $s_{\max}$ with parsimonious representation used for retaining three diffusion coordinates. While smaller $s_{\max}$ values generally yielded less accurate surrogates, $s_{\max} = 15$ exhibited the largest validation errors for all p. Larger manifold dimensions led to increased mean errors and variability. Based on this analysis, we used $s_{\max} = 25$ to obtain Sobol' indices on the manifold.

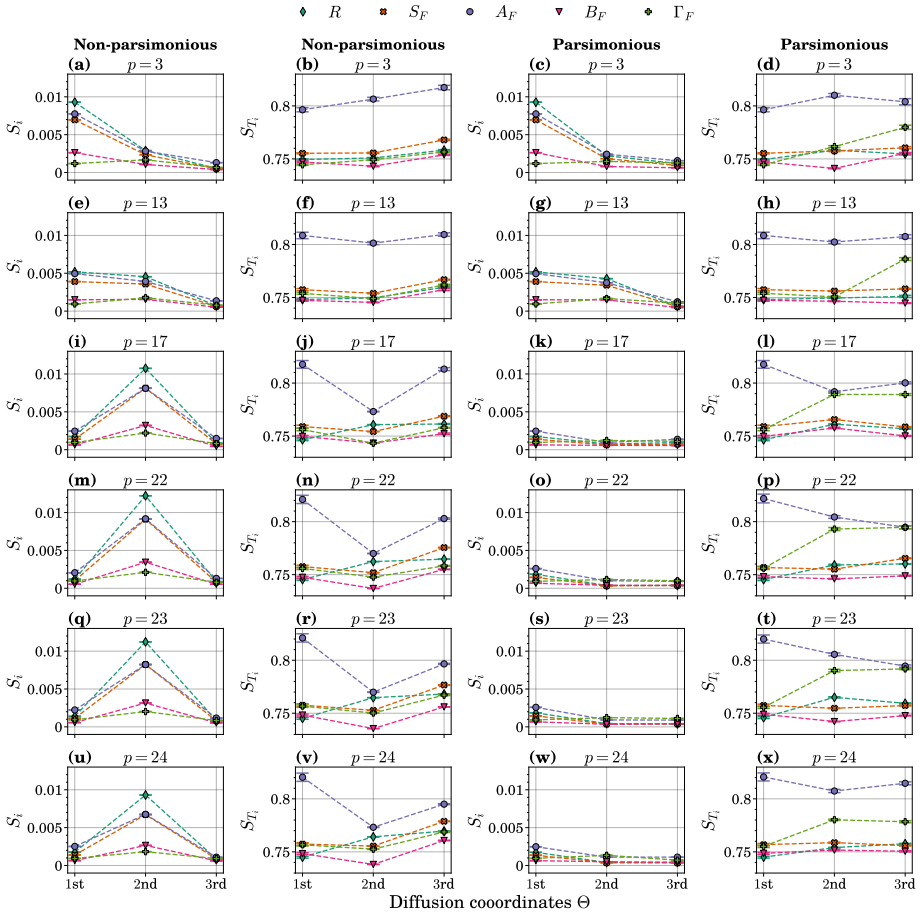

Fig. 5. Estimates of first- and total-order sensitivity indices, S_i and S_{T_i}, respectively, of five uncertain input parameters (see SI Section B.3 Tab. 2) for five model output measures, obtained from the GSA framework using GDMaps PCE. Error bars indicate variance across 20 runs.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

Fig. 5 presents first- and total-order sensitivity indices, averaged over twenty runs, from PCE using GDMaps with six Grassmannian dimensions, $p = \{3, 13, 17, 22, 23, 24\}$, and both non-parsimonious and parsimonious representations to keep three diffusion coordinates. As observed in Figs. 5a to 5d and Figs. 5e to 5h, S_i and S_{T_i} are comparable across representations for $p = 3$ and $p = 13$. This is due to GDMaps with lower p values resulting in larger initial eigenvalues, leading to parsimonious representation selecting diffusion coordinates that correspond to longer timescale dynamics more frequently. The main difference between the two implementations can be seen for $p = \{17, 22, 23, 24\}$, especially in the main effect indices, S_i. Larger p values provide more detailed data representation on the manifold, while non-parsimonious selection results in diffusion coordinates capturing the longest timescale dynamics on the manifold, possibly related to individual parameter contributions to model output variance. Overall, individual parameter contributions to output variance are smaller compared to interaction effects, consistent with results using conventional methods for obtaining Sobol' indices, which can be found in SI Fig. C.4.

An interesting aspect of the proposed framework is the impact of different maximum polynomial degrees $s_{\max}$ used in PCE basis construction on resulting Sobol' indices. Increasing $s_{\max}$ allowed for more accurate model output reconstruction (reducing validation error) and higher total-order sensitivity indices for all parameters and Grassmannian dimensions (see SI Fig. C.10). Lower $s_{\max}$ values resulted in higher first-order sensitivity indices, especially for the first diffusion coordinates and lower Grassmannian dimensions (see SI Fig. C.9).

5 Discussion and Conclusions

The proposed method for parametric GSA utilises a manifold learning-based approach to construct PCE emulators on lower-dimensional manifolds for high-dimensional problems with significant interaction effects. Unlike traditional methods, this technique enables a more general estimation of parametric sensitivities by aggregating entire trajectories of multiple model responses. Using an oscillating model example, we demonstrated that traditional Sobol' index estimation failed to provide a definitive answer to what parameter is relatively more important and whether the variance in model output is influenced by main or interaction effects. Conversely, the GSA method employing GDMaps PCE successfully revealed clear relations between parameters and their relative influence on the output variance.

Characterised by non-linearity, ABMs are suitable candidates for the GSA approach using PCE on the Grassmann manifold. In this paper, we applied the method to a large-scale spatial ABM of epidemic dynamics and captured strong interaction effects of uncertain parameters on the variance of multiple aggregated model responses. We also investigated the influence of hyper-parameters, such as the dimension of the Grassmann manifold and maximal polynomial degree and two approaches for retaining the desired number of diffusion coordinates, parsimonious and non-parsimonious, on sensitivity measures. Lower Grassmannian

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

dimensions yielded higher main effect indices due to a more coarse-grained data representation. Non-parsimonious implementation produced larger first-order indices, resulting from longer timescales of diffusion coordinates corresponding to larger eigenvalues. Additionally, a higher maximal polynomial degree was attributed to a smaller validation error, as expected, which was mainly caused by the resolution of the interactions between parameters, leading to larger total-effect indices.

A simulation outcome or trajectory is a time series of state variables with dimension kN, where N is the number of agents and k is each agent's state variables. As typical agent-based simulations may involve a considerable number of agents, these time series can become exceedingly high-dimensional. As part of our future work, we aim to leverage the power of GDMaps to reduce the complexity of simulation trajectories at the micro level. By doing so, we can explore the impact of parameter sensitivity on the system's dynamic modes, key long-lasting states, transitional pathways, and essential degrees of freedom. Furthermore, we plan to address a limitation in the current implementation of not following general advice to perform over-sampling for the regression used in calculating PCE coefficients [9]. Model selection, like Least Angle Regression, should be added to the implementation to circumvent this issue.

In conclusion, the capabilities of the GSA framework utilising GDMaps PCE to aggregate entire trajectories of multiple model outputs and capture different timescales and degrees of structural complexity satisfy all the "good" properties of a global sensitivity measure extensively discussed in [21], providing a more comprehensive estimation of parameter sensitivities. This methodology is expected to open new avenues for ABM practitioners and Complexity Science scholars to deepen their understanding of systems exhibiting complex spatiotemporal dynamics.

Supporting Information The Supporting Information can be found at Valentina Bazyleva, Victoria Garibay, & Debraj Roy. (2023). Supporting Information: Global Sensitivity Analysis using Polynomial Chaos Expansion on the Grassmann Manifold. Zenodo. https://doi.org/10.5281/zenodo.7852159.

Data and Code Availability The code used in generating data for testing the methodology proposed in this study can be found at https://github.com/bazvalya/GSA_using_GDMaps_PCE. The output data of the DeepABM-COVID simulations is at https://figshare.com/articles/dataset/output_data_zip/22216921.

Acknowledgements This research was conducted with support from the Dutch Research Council (NWO) under contract 27020G08, titled "Computing societal dynamics of climate change adaptation in cities".

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

References

1. Baustert, P., Benetto, E.: Uncertainty analysis in agent-based modelling and consequential life cycle assessment coupled models: a critical review. Journal of Cleaner Production **156**, 378–394 (2017)
2. Cacuci, D.G.: Sensitivity & Uncertainty Analysis, Volume 1. Chapman and Hall/CRC, 0 edn. (May 2003). https://doi.org/10.1201/9780203498798
3. Chopra, A., Gel, E., Subramanian, J., Krishnamurthy, B., Romero-Brufau, S., Pasupathy, K.S., Kingsley, T.C., Raskar, R.: Deepabm: Scalable, efficient and differentiable agent-based simulations via graph neural networks (2021). https://doi.org/10.48550/ARXIV.2110.04421
4. Colas, F., Gauchi, J.P., Villerd, J., Colbach, N.: Simplifying a complex computer model: sensitivity analysis and metamodelling of an 3d individual-based crop-weed canopy model. Ecological Modelling **454**, 109607 (2021)
5. Crestaux, T., Le Maıtre, O., Martinez, J.M.: Polynomial chaos expansion for sensitivity analysis. Reliability Engineering & System Safety **94**(7), 1161–1172 (2009)
6. Dos Santos, K.R., Giovanis, D.G., Shields, M.D.: Grassmannian diffusion maps–based dimension reduction and classification for high-dimensional data. SIAM Journal on Scientific Computing **44**(2), B250–B274 (2022)
7. Dsilva, C.J., Talmon, R., Coifman, R.R., Kevrekidis, I.G.: Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. Applied and Computational Harmonic Analysis **44**(3), 759–773 (2018)
8. Edeling, W., Arabnejad, H., Sinclair, R., Suleimenova, D., Gopalakrishnan, K., Bosak, B., Groen, D., Mahmood, I., Crommelin, D., Coveney, P.V.: The impact of uncertainty on predictions of the covidsim epidemiological code. Nature Computational Science **1**(2), 128–135 (2021)
9. Hosder, S., Walters, R., Balch, M.: Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. In: 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. p. 1939 (2007)
10. Hu, Y.: Agent-based models to couple natural and human systems for watershed management analysis. Ph.D. thesis, University of Illinois at Urbana-Champaign (2016)
11. Kalogeris, I., Papadopoulos, V.: Diffusion maps-based surrogate modeling: An alternative machine learning approach. International Journal for Numerical Methods in Engineering **121**(4), 602–620 (2020)
12. Kalogeris, I., Papadopoulos, V.: Diffusion maps-aided neural networks for the solution of parametrized pdes. Computer Methods in Applied Mechanics and Engineering **376**, 113568 (2021)
13. Konakli, K., Sudret, B.: Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions. Journal of Computational Physics **321**, 1144–1169 (2016)
14. Kontolati, K., Loukrezis, D., dos Santos, K.R., Giovanis, D.G., Shields, M.D.: Manifold learning-based polynomial chaos expansions for high-dimensional surrogate models. International Journal for Uncertainty Quantification **12**(4) (2022)
15. Kucherenko, S., Song, S.: Different numerical estimators for main effect global sensitivity indices. Reliability Engineering & System Safety **165**, 222–238 (2017)
16. Ligmann-Zielinska, A., Sun, L.: Applying time-dependent variance-based global sensitivity analysis to represent the dynamics of an agent-based model of land use change. International Journal of Geographical Information Science **24**(12), 1829–1850 (Nov 2010). https://doi.org/10.1080/13658816.2010.490533

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46

17. Lüthen, N., Marelli, S., Sudret, B.: Sparse polynomial chaos expansions: Literature survey and benchmark. SIAM/ASA Journal on Uncertainty Quantification **9**(2), 593–649 (2021)
18. Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E.: A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of theoretical biology **254**(1), 178–196 (2008)
19. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **42**(1), 55–61 (2000)
20. Owen, A.B.: Better estimation of small sobol'sensitivity indices. ACM Transactions on Modeling and Computer Simulation (TOMACS) **23**(2), 1–17 (2013)
21. Pianosi, F., Wagener, T.: A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling Software **67**, 1–11 (2015). https://doi.org/https://doi.org/10.1016/j.envsoft.2015.01.004
22. Richiardi, M.G., Leombruni, R., Saam, N.J., Sonnessa, M.: A common protocol for agent-based social simulation. Journal of artificial societies and social simulation **9** (2006)
23. Saltelli, A. (ed.): Global sensitivity analysis: the primer. John Wiley, Chichester, England ; Hoboken, NJ (2008), oCLC: ocn180852094
24. Saltelli, A.: Making best use of model evaluations to compute sensitivity indices. Computer physics communications **145**(2), 280–297 (2002)
25. dos Santos, K.R., Giovanis, D.G., Kontolati, K., Loukrezis, D., Shields, M.D.: Grassmannian diffusion maps based surrogate modeling via geometric harmonics. International Journal for Numerical Methods in Engineering **123**(15), 3507–3529 (2022)
26. Smith, R.C.: Uncertainty quantification: theory, implementation, and applications, vol. 12. Siam (2013)
27. Sobol', I.M., Myshetskaya, E.: Monte carlo estimators for small sensitivity indices **13**(5-6), 455–465 (2008). https://doi.org/doi:10.1515/mcma.2007.023
28. Sobol', I.M.: On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki **7**(4), 784–802 (1967)
29. Sobol, I.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation **55**(1-3), 271–280 (Feb 2001)
30. Sudret, B.: Global sensitivity analysis using polynomial chaos expansions. Reliability engineering & system safety **93**(7), 964–979 (2008)
31. Tang, W., Jia, M.: Global sensitivity analysis of a large agent-based model of spatial opinion exchange: A heterogeneous multi-gpu acceleration approach. Annals of the Association of American Geographers **104**(3), 485–509 (2014)
32. Tarantola, S., Gatelli, D., Kucherenko, S., Mauntz, W., et al.: Estimating the approximation error when fixing unessential factors in global sensitivity analysis. Reliability engineering & system safety **92**(7), 957–960 (2007)
33. Wainwright, H.M., Finsterle, S., Jung, Y., Zhou, Q., Birkholzer, J.T.: Making sense of global sensitivity analyses. Computers & Geosciences **65**, 84–94 (Apr 2014). https://doi.org/10.1016/j.cageo.2013.06.006

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_46